SPARK Pairs

method map(row)

    list <- []

    movie_list <- split the row

    sort movie_list

    for all movie_i in movie_list **do**

        for all movie_j > movie_i in movie_list **do**

            list.append(((movie_i, movie_j), 1)

    return list

method load_movie_names()

    read movies.csv file and store the movie names into a global dictionary

    where movie_id is key and movie_name is the value

method main()

    read text file as RDD

    RDD.map(split row into user_id, movie_id, rating)

    RDD.filter(rows with rating >= 4.0)

    RDD.map(row <- (user_id, movie_id))

    RDD.map(for each user_id concatenate all movie_id's)

    RDD.map(row <- concatenated movie_id's)

    RDD.map(map)

    RDD.reduceByKey(sum the counts for each pair)

    RDD.filter(pairs with count > threshold)

    global_movie_names <- load_movie_names()

    RDD.map(row <- movie_name_1, movie_name_2, count)

    save RDD as text file